Assessing the Predictive Influence of Human, Road Conditions, and Vehicle Factors on Accident Severity in Victoria

Tazeen Atif
tazeen.atif20@gmail.com

Nikhil Gaba
nikhilkumargaba@gmail.com

Shwethan Reddy Potu
potushwethan2005@gmail.com

Executive Summary

The report investigates the extent to which human, road conditions, and vehicle-related factors can predict accident severity on Victorian roads. Using the Victorian Road Crash dataset, the analysis focuses on a range of predetermined features, including seatbelt usage, age group, road surface conditions, atmospheric conditions, node type, fuel type, and vehicle type. The primary objective is to uncover which of these factors are most strongly associated with the severity of road accidents, and to assess their predictive value using statistical and machine learning techniques.

The data undergoes preprocessing techniques, including mode imputation for missing values, one-hot encoding for categorical variables, and discretisation of age. Heatmaps are used to visualise the distribution of accident severity across subcategories, highlighting trends such as higher fatal accidents amount elderly individuals, increased severity for unbelted occupants, and more severe outcomes for heavy freight vehicles and low-visibility road conditions.

Mutual information scores are calculated to quantify the dependency between each factor and accident severity. Seatbelt use and age group (particularly the elderly 65+) show the highest Mutual Information scores, potentially indicating their strongest predictive power. Supervised learning models (SLM), including Decision Tree and K-Nearest Neighbours classifiers, are implemented to test the practical predictability of these factors. However, both models primarily predict the majority class and struggle to identify minority severity outcomes, indicating limitations in the feature set and the impact of class imbalance.

Key recommendations include targeted road safety campaigns focused on seat belt compliance, age-specific interventions, and improved regulations for freight vehicle operations. These findings provide insights to reduce the severity of road accidents and improve the safety on Victorian Roads.

Introduction

The research question for this report is: **To what extent are predetermined factors, including seat belt usage, age, road surface conditions, atmospheric conditions, node type, fuel and vehicle types, predictive of accident severity on Victorian roads?**

This report investigates how specific, predetermined factors - seatbelt use, age, road surface conditions, atmospheric conditions, node type, fuel and vehicle features - contribute to the report's dependent variable - accident severity - on Victorian roads. The data used for this investigation is 'Victorian Road Crash Data' from Victoria Police Reports , comprising nine datasets with exhaustive data from each accident, identified by a unique accident number. This data is cleaned and preprocessed, and visualised to see past trends and patterns. Then, it is used in exploratory analysis predictive modelling techniques, to assess the relative predictive influence of each factor. Mutual information scores are used to measure non-linear dependencies between factors and accident severity. Classification models, such as Decision Trees and K-Nearest Neighbours (KNN) are implemented to evaluate the practical predictive influence of the factors.

Methodology

Preprocessing - Imputation

Table 1: Proportion Unknown Values of Total Records for Each Predetermined Feature

Feature	Age Group	Helmet Belt Worn	Road Surface	Atmospheric Condition	Node Type	Fuel Type	Vehicle Type
Proportion Unknown Values	3.17%	27.08%	6.35%	9.97%	<0.01%	2.44%	0%

Before we consider key methods of imputation, it's important to assess whether values are missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR). Figure 1 demonstrates that from the accidents' road surface conditions that were missing, 76.7% of these accidents also had atmospheric conditions missing. Comparing this figure to 9.97% missing values in all atmospheric condition data (Table 1), it can be observed that it is more likely for an accident's atmospheric condition to be missing, given that the accident's road surface condition is also missing. Similarly, the proportion of missing values in age groups subset given road condition data is unknown is 5.9% (Figure 1), which is larger than the 3.17% of missing values in the full data (Table 1). Therefore, this data is known as MAR since the probability of missing values in a column is not independent of other columns' missing data. For the accidents where the road surface condition was missing, the fuel type data is missing only 0.2% of the time, suggesting that the data may be missing completely at random (MCAR). Likewise, there does not appear to be any missing data in node types, however, since there are only five missing values, the nature of the data in terms of whether it is MCAR or MAR is negligible.

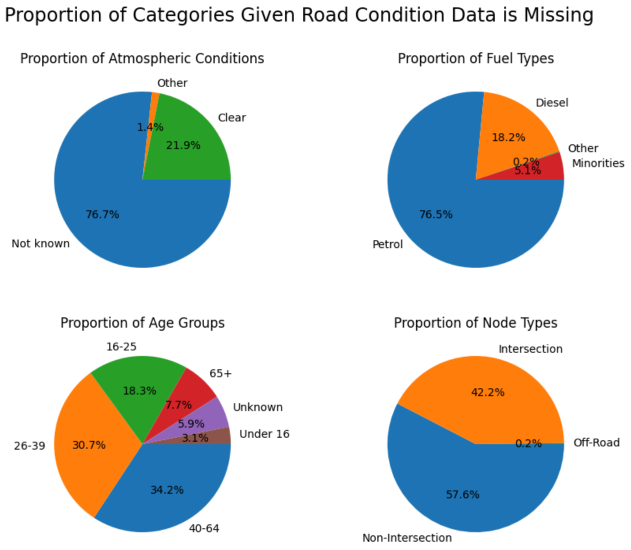

Figure 1: Proportion of Categories for Each Feature For Subset of Missing Data for Road Surface Condition

Since all data is either MCAR or MAR, it is appropriate to use simple imputation methods as preprocessing steps to ensure that useful data is not lost. Generally, imputation of the mode category was used, since the mode category was quite frequent, with one clear majority category clearly present in the data (see Figure 2 and 3), which was also coincidentally the median, suitable for categorical data. However, there were limitations considered including not imputing a significant number of rows, such that the resulting data distribution is not dissimilar to before, as to prevent bias. For our predetermined features, Age Group, Road Surface Condition, Atmospheric Condition and Node Type, Fuel Type have small proportions of missing values, 3.17%, 6.35%, 9.97%, <0.01% and 2.44% respectively (Table 1), which did not skew the resulting distribution significantly (see Figures 2 and 3). It is worth noting that Age Group did not have a 'clear majority,' however, imputing the small proportion of missing values as the most frequent category did not skew the resulting distribution significantly. However, Helmet/Belt Worn had a significantly impactful proportion of missing values, 27.08%, which would have significantly skewed the distribution or distorted the underlying class distributions, if handled through imputation of mode.

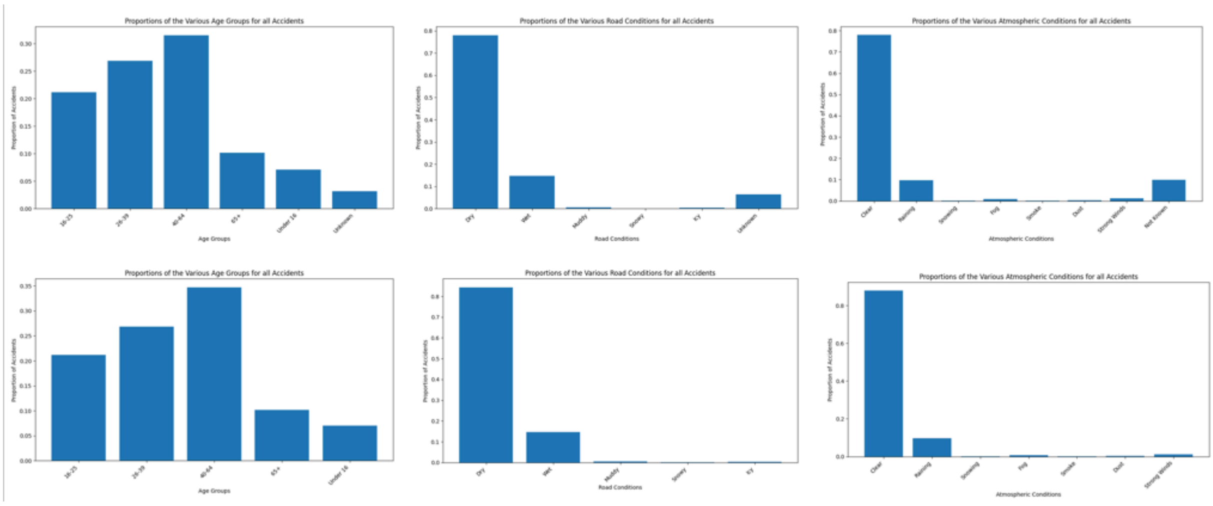

Figure 2: Distribution of Categories for the Various Predetermined Features Before (Above) and After (Below) Imputation

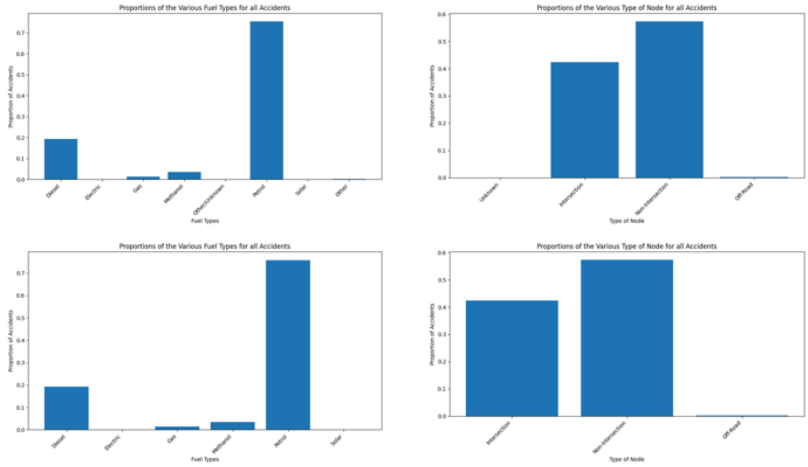

Figure 3: Distribution of Categories for the Various Predetermined Features Before (Above) and After (Below) Imputation

Imputing was largely preferred to the alternative of leaving the 'Unknown' values as a separate informative category. This was because there were generally low percentages of missing values for all categories and imputing these to a clear majority category seemed a reasonable assumption as the resulting distribution was similar to the pre-imputed distribution. However, the predetermined features, vehicle type and whether or not helmets and seat belts were worn were not imputed, for differing reasons. Firstly, helmet belts worn had a very large proportion of unknown values, approximately 27% (from Table 1). Imputing all this data could result in significant bias to the results, so the unknown category was unchanged and left on its own as an informative category. Vehicle type, on the other hand, had no missing values, so it was natural that no imputation took place.

Encoding - One-hot encoding

Another key preprocessing technique used was one-hot encoding of categorical variables, such as seat belt usage, road surface and atmospheric conditions, fuel type and vehicle type descriptor to convert the required data into a numerical format. Label encoding was not used as this preprocessing step introduces notions of arbitrary distance and ordinal relationships, where a category may be assigned a larger numeric value and (incorrectly) be perceived as greater than another category. Hence, one-hot encoding was preferred as an alternative.

One-hot encoding converts all entries for a feature into a vector of length k, where k is the number of unique categories for a feature. A unique characteristic of this is that exactly one entry of each vector is one, indicating the presence of this category, and all others are zero, indicating the absence of the remaining categories. This encoding technique was preferred ahead of label encoding, as it provides a solution to the ordinal relationships described earlier.

Transformation - Discretisation

In essence, Age, one of the predetermined factors of this research, is of a continuous data format, and it should be discretized into age groups to ensure consistency with the other predetermined features. In the person dataset, the column 'AGE_GROUP' had unordered age ranges, with no categorisation formula applied. To capture the potential non-linear effects of this factor, discretized age was used in the data visualisation and learning model. The technique also helps reduce noise and overfitting in the decision tree model used later on.

Merging Datasets

As part of the preprocessing stage of the research, it was essential to merge the required data subsets, consisting of the predetermined features as well as the accident severity in one data object for further analysis, such as correlation, training and test data splits and supervised learning models. Merging was performed using a common identifier, ACCIDENT_NO, to ensure consistency across records.

Data exploration and analysis

Heatmaps - Proportion of Accident Severity for each predetermined factor

The heatmaps in Figures 4-10 visualise the proportions of accident severity levels across the different categorical features chosen to be the specific, predetermined factors of this report, in reference to assessing their predictive influence for accident severity. Each heatmap breaks down accident outcomes - classified as fatal, serious, minor injuries or no injuries - by the given factors. This visual approach helps identify which subgroups are more prone to severe outcomes and reveals potential risk factors influencing accident severity.

It can be observed from Figure 4 that accidents in snowy conditions have significantly higher 'minor injury' accidents and significantly lower 'serious injury' and fatal accidents compared to other road surface conditions, such as wet, dry and icy conditions. 81% of accidents reported in snowy conditions were non-serious injuries, compared to about 60% for other road surface conditions. Likewise, only 18.7% of such accidents were serious and 0.3% were fatal, compared to about 35% serious and 2% fatal for the other conditions. A potential explanation of this could be that drivers tend to drive much slower in snowy conditions recognising the potential for vehicles to swerve in such conditions. Slower speeds may reduce the likelihood of high-impact accidents, with very severe outcomes. With regards to other road conditions, there was not a significant difference in accident severity distributions.

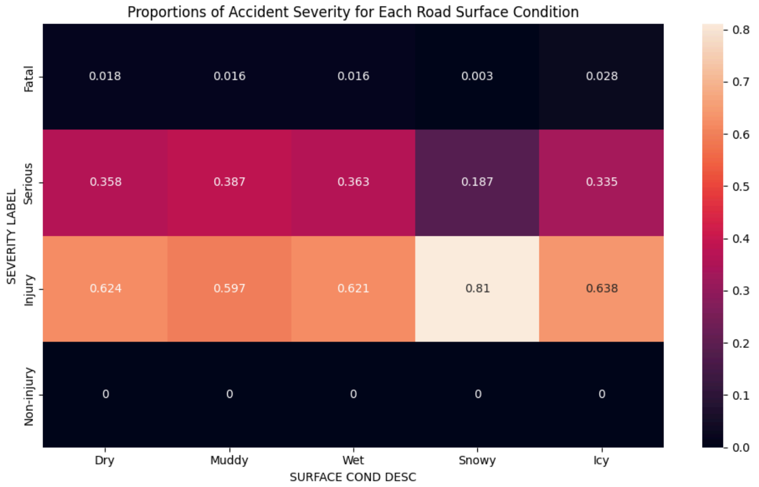

Figure 4: Proportion of Accident Severity Labels for Each Road Surface Condition

The proportions of accident severity for each atmospheric condition heatmap (see Figure 5) shows a higher proportion of accidents with serious injuries and a lower proportion of accidents with minor injuries for rarer atmospheric conditions such as fog, dense smoke or dust, impacting visibility. Accidents reported in such atmospheric conditions have about 43% of serious injuries and 55% of minor injuries, compared to 35% of serious accidents and 62% of not severe accidents on average for the other weather conditions. This suggests that reduced visibility may impact reaction times of drivers, which may lead to increased accident severity outcomes. Moreover, it should also be noted that drivers are more accustomed to driving in clear, raining and other such weather conditions, hence, more adept at reacting to conditions at hand.

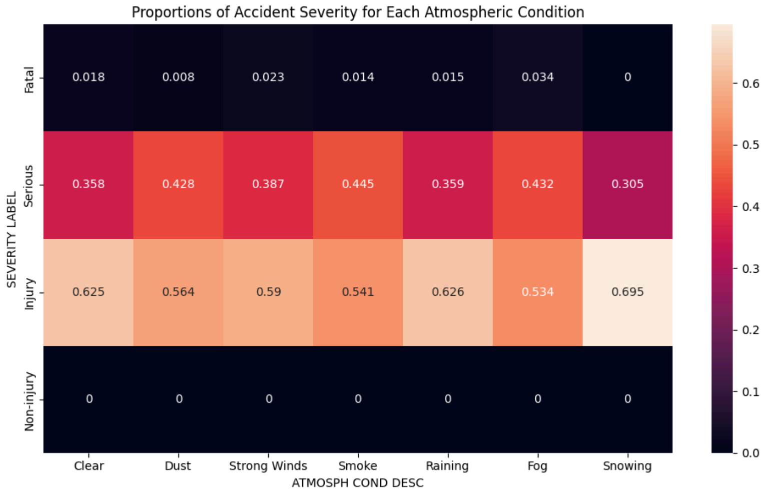

Figure 5: Proportion of Accident Severity Labels for Each Atmospheric Condition

From Figure 6, it can be observed that accidents that occur at non-intersections, such as freeways, tend to have more serious consequences in terms of accident severity. 2.3% and 37.2% of such accidents are fatal and serious respectively, compared to 1.2% and 34.3% of accidents at intersections and 0.8% and 34.4% off the road. Naturally, there was a lower proportion of less severe accidents for non-intersection nodes (see Figure 5). This can potentially be explained with higher average speeds on freeways, and other carriageways, where drivers do not expect to have to brake, resulting in greater accident severity, as depicted in the heatmap in Figure 6.

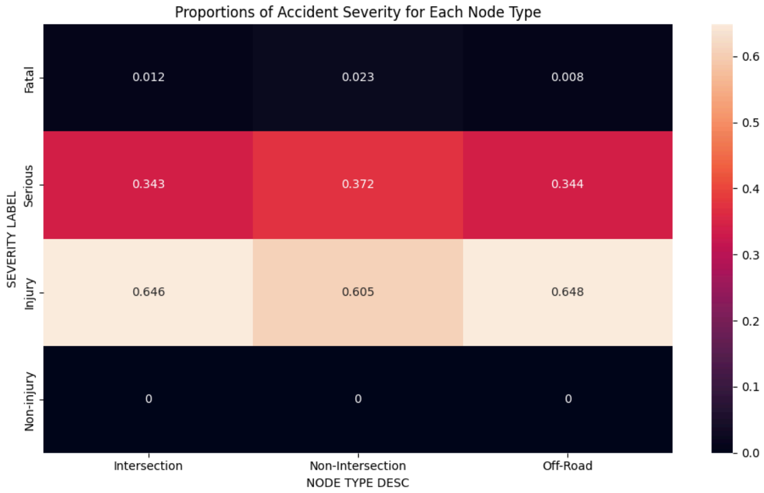

Figure 6: Proportion of Accident Severity Labels for Each Node Type

Figure 7 is a heatmap displaying the proportions of accident severity for cases where a seatbelt was worn or not worn. It can be observed that accidents where individuals were not wearing seatbelts resulted in significantly higher proportions of fatal and serious injuries, compared to those who were wearing seatbelts. In terms of 'Seatbelt Not Worn,' 6.4% of accidents resulted in fatalities, and 46% resulted in serious injuries, while 1.6% of 'Seatbelt Worn' accidents resulted in fatalities, and 34.4% resulted in serious injuries. Conversely, 64% of accidents with seatbelt users and 47.5% of accidents with non-seatbelt users had less severe injuries. Non-injury cases were not observed in the dataset, suggesting all recorded accidents resulted in some form of injury. These results display the trend of wearing seatbelts greatly reducing the likelihood of sustaining fatal or serious injuries in an accident. Since seatbelts restrain occupants during a collision, and hence, reduce the force of impact and prevent ejection from the vehicle.

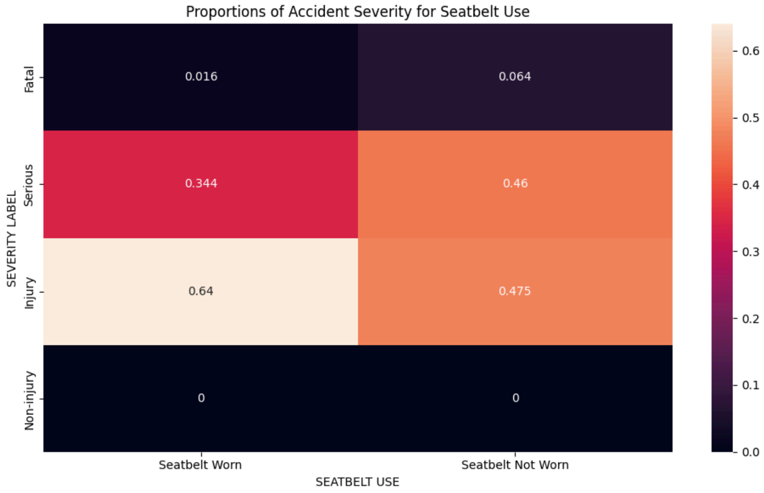

Figure 7: Proportion of Accident Severity Labels for Seatbelt Use

The heatmap observations in Figure 8 are based on the proportions of accident severity varying across age groups. The 65+ age group experiences the highest proportion of fatal and serious injuries, with 2.9% of accidents involving this age group resulting in fatalities, whereas 1.6-1.7% were recorded for fatal accidents involving younger age groups. Similarly, 44.2% of accidents in the 65+ age group were serious, compared to the younger age groups ranging between 33.7% to 36.6% in terms of serious severity. In contrast, the 26-39 age group has the highest proportion of less-severe injuries, at 64.8%, while the 65+ age group has the lowest proportion in this category, at 52.9%. The proportion of non-injury cases remained at 0% across all age groups, indicating that all recorded incidents involved some form of injury. These results suggest that older age groups are more vulnerable to severe accidents, fatal or serious. A possible explanation may be due to old-age features, such as slower reaction time and confusion in accidents. In terms of injuries, physical weakness/frailty and slow recovery from trauma may also be responsible for these significant results.

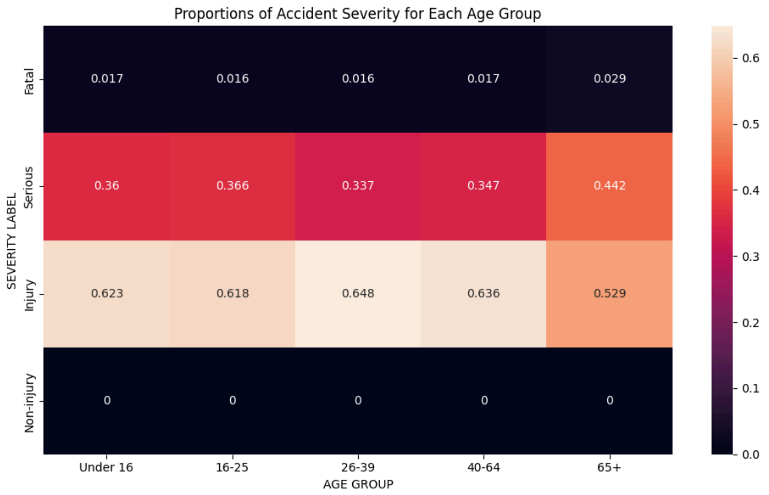

Figure 8: Proportion of Accident Severity Labels for Each Age Group

Figure 9 presents a heatmap illustrating the proportions of accident severity across different fuel types. Most fuel types, including Diesel, Electric, Gas, Methanol, Petrol, and Solar show relatively consistent distributions of accident severity. For these fuels, fatal accidents remain low, ranging from 1.3% to 3.1% for the most part apart from solar fueled vehicles, and serious injuries account for roughly one-third of all cases, between 33.3% and 36.7%, not considering solar fueled vehicles. The majority of incidents involving these vehicles resulted in injuries, with proportions between 60.7% and 65.2%. However, solar-powered vehicles stand out dramatically, with 71.4% of accidents resulting in fatalities and 28.6% leading to injury, and no recorded cases of serious injury. This skewed distribution is likely influenced by a very small number of solar vehicle accidents, meaning even a single fatality heavily shifts the proportions. Across all fuel types, the non-injury row remains at 0%, indicating that every recorded accident led to some form of injury. Overall, this analysis suggests that conventional fuel types are associated with more moderate and predictable severity outcomes, while extreme values, such as those for Solar-powered vehicles may be due to rare or outlier cases. These findings reinforce the reliability of safety data among mainstream fuel technologies and highlight the importance of careful interpretation when dealing with underrepresented categories.

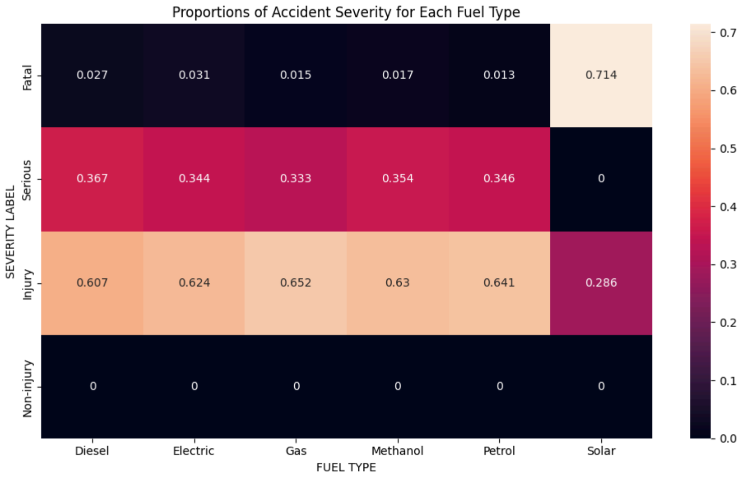

Figure 9: Proportion of Accident Severity Labels for Each Fuel Type

Figure 10 is a heatmap showing the distribution of accident severity across various vehicle types. Most vehicle categories show a predominant proportion of injury-level accidents, with Light Commercial Vehicles, Cars, and Motorcycles reporting injury proportions between 61.5% and 64.4%. In contrast, vehicles like Rigid Trucks and Prime Mover > 1 Trailers demonstrate much higher proportions of serious injuries, at 74% and 75.6%, respectively. These values are notably higher than the serious injury rates observed in more common vehicle types, which typically range between 33% and 52%. The fatality proportions remain relatively low across the board, generally below 5.5%, with the exception of Prime Mover B-Double and Prime Mover B-Triple vehicles, which exhibit elevated fatality rates of 16% and 6.6%, respectively—suggesting that larger commercial vehicles may be involved in more severe accidents

Parked Trailers stand out in the dataset with a 100% serious injury rate, which is likely a result of a very small sample size and should be interpreted with caution. Across all vehicle types, the non-injury category remains at 0%, indicating that all recorded incidents involving these vehicles resulted in at least some form of physical injury. Overall, the analysis reveals a pattern where heavier or commercial-use vehicles tend to be associated with higher severity outcomes, especially serious injuries and fatalities, while lighter or passenger vehicles are more often linked with less severe injuries. These trends underscore the need for stricter safety measures and design regulations specifically for freight and industrial vehicle categories, given their potential to cause disproportionately severe harm in the event of an accident.

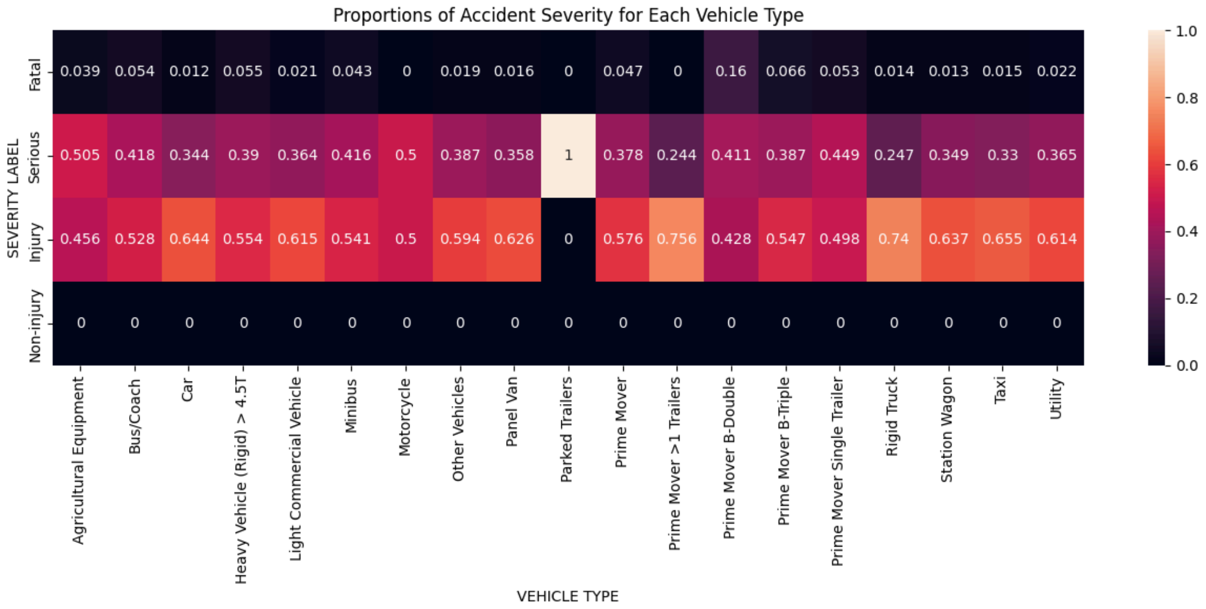

Figure 10: Proportion of Accident Severity Labels for Each Vehicle Type

MI Scores for predetermined factors

Figure 11 displays the Mutual Information scores for each subcategory of the specific, predetermined factors of this report, in relation to accident severity. Mutual information measures the dependency between two variables - in this case, how much information knowing a specific subcategory provides about the severity of an accident. Higher values indicate stronger associations, which can be observed through the age group 65+ showing the highest mutual information scores, and Non-Intersection and Intersection node types being approximately second and third highest. This suggests they are the most informative predictors of accident severity. Older age groups, especially those aged 65+, and non-intersection road types also show moderate dependency with accident severity, highlighting their potential influence on severe accidents. Conversely, weather-related subcategories (such as fog, snow, and strong winds) and road surface conditions (like wet or snowy roads) have relatively low mutual information scores. While these factors may visually correlate with severity in heatmaps, they do not contribute significantly to the prediction of severity when considered in isolation from other features.

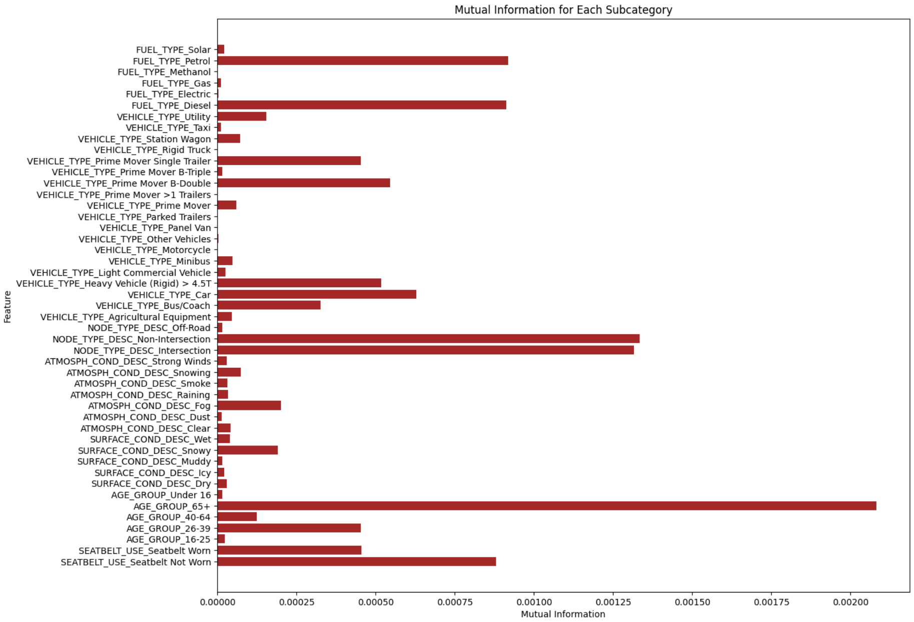

Figure 11: Correlation of Each Category from Each Feature as measured by Mutual Information

Classification Model - Decision Tree

For the decision tree classifier based on the three highest correlated subfeatures, the model had a respectable overall accuracy of 0.632 and a weighted average F1-score of 0.490 (Table 2). From Figure 12, it is clear that the decision tree always predicts ‘Other Injury’, regardless of the independent variables chosen from feature selection. ‘Other Injury’ is the majority category within the test set, comprising 100,360 accidents of the total 158,725 accidents in the test set (Table 2), evaluating to a proportion of 63.23% of accidents classified within this category. This suggests a very poor efficacy of the model since variation in the three highest correlated sub features do not explain accident severity well in the decision tree classifier.

Furthermore, from Table 2, we see that the recall is 1 for ‘Other Injury’, indicating that when the true accident severity is ‘Other Injury’, it always correctly predicts it. This is a trivial result since the classifier always predicts this severity label. Naturally, the recall of all other severity labels, the proportion of time the classifier correctly predicts that class label is 0, since it never predicts any other label. Another observation is that the precision for ‘Other Injury’, the proportion of correct predictions out of all predictions for that label, is equivalent to the total proportion of ‘Other Injury’ accidents in the test set. All factors combined, this model’s efficacy is extremely low, evidenced by the 0.194 macro average F1-score (Table 2).

Table 2: Evaluation Metrics for the Decision Tree Classifier

	Precision	**Recall**	**f1-score**	**support**
Fatal	0.000	0.000	0.000	2539
Serious Injury	0.000	0.000	0.000	55824
Other Injury	0.632	1.000	0.775	100360
Non-injury	0.000	0.000	0.000	2
Accuracy			0.632	158725
Macro Avg	0.158	0.250	0.194	158725
Weighted Avg	0.400	0.632	0.490	158725

From the confusion matrix in Figure 12, it is worth noting that when 'Other Injury' is predicted, there are 55824 (35.17%) serious accidents in the test set (incorrectly) predicted as other injuries, indicating a moderate precision.

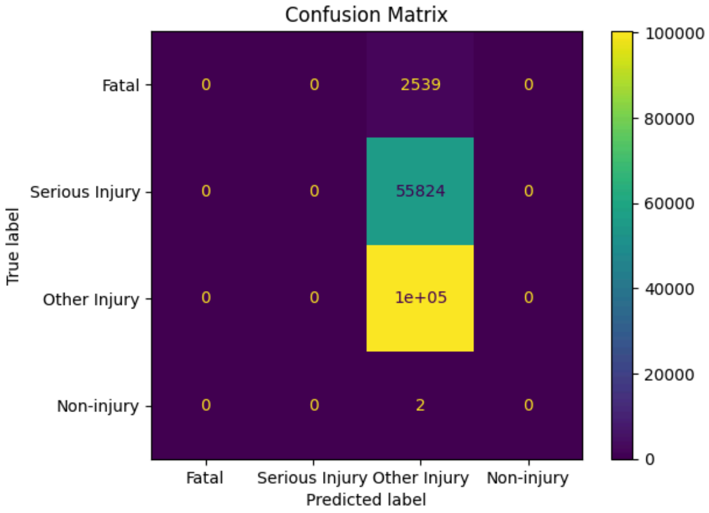

Figure 12: Confusion Matrix of Decision Tree Classifier based off Three Highest Correlated Features

Classification Model - K-Nearest Neighbours

The k-Nearest Neighbours (k-NN) classified, based on the top three correlated subfeatures, achieved an overall accuracy of 0.625, and a weighted average F1-score of 0.519 (Table 3). This accuracy value appears acceptable at an initial glance, but it aligns extremely closely with the prevalence of the majority class - '3: Other Injury.' The class accounts for 100 360 of the total 158 725 accidents, approximately 63.23% of the dataset. This close match in accuracy and class distribution indicates that the k-NN model performs just marginally better than always predicting the majority category.

Observations from Table 3 relay that the recall for 'Other Injury' is 0.956, which means that the k-NN model is highly effective at recognising this majority class. Meanwhile, the model's performance for all the other classes - Fatal, Serious and Non-Injury, is relatively poor. The recall for 'Serious Injury' is just 0.058, and the other two have values of 0.000, meaning that the model fails completely in identifying these classes. Precision values for these minority classes are slightly higher, but are still relatively low due to the infrequency with which the model assigns these labels. The resulting F1-scores for the minority classes are also extremely low, further underscoring the model's lack of sensitivity to less frequent, yet vital, severity levels. Ultimately, the k-NN classifier exhibits a strong bias towards the dominant class - Other Injury. Hence, it offers little practical use in classifying rare, but severe, outcomes.

Table 3: Evaluation Metrics for the K-Nearest Neighbour Classifier

	Precision	**Recall**	**f1-score**	**support**
Fatal	0.000	0.000	0.000	2539
Serious Injury	0.418	0.058	0.102	55824
Other Injury	0.636	0.956	0.764	100360
Non-injury	0.000	0.000	0.000	2
Accuracy			0.625	158725
Macro Avg	0.263	0.254	0.216	158725
Weighted Avg	0.549	0.625	0.519	158725

From the confusion matrix in Figure 13, it is worth noting that when 'Other Injury' is predicted, there are 52583 (33.13%) serious accidents in the test set (incorrectly) predicted as other injuries, indicating a moderate precision.

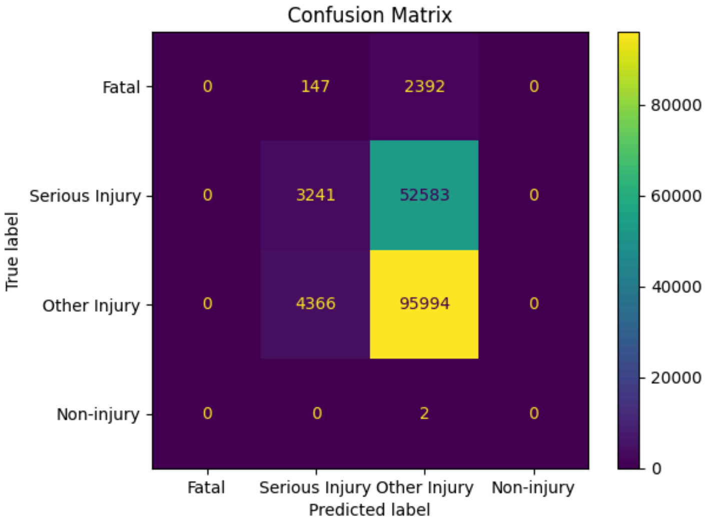

Figure 13: Confusion Matrix of KNN Classifier based off Three Highest Correlated Features

Discussion and interpretation

MI Correlation for each predetermined factor

We can conduct correlation analysis as to what categories of features are most associated with accident severity, however, there are some limitations to consider before conducting the analysis. We shall use mutual information scores to detect any non-linear dependencies between each subcategory and accident severity, especially since our pre-processed data is of a discrete numerical form, and not continuous numerical, yielding Pearson's correlation coefficient inapplicable.

From Figure 11, we can start to conjecture as to what features correlate most with accident severity, and start to select the most informative features for supervised learning models to investigate as to how predictive features are. None of the mutual information scores are exceptionally high, with the highest at 0.0026 for wearing a seatbelt. Since accident severity likely depends on countless factors, with no one factor solely determining accident severity, this explains why the very low correlations, despite having theoretical evidence against this. The features most correlated with accident severity are seatbelt usage, whether seat belts were worn or not, shielding the impact in the event of a crash, and the elderly (65 years and above). This correlation aligns with the premeditated outlook in the heatmap in Figure 7, since we see that wearing a seatbelt is much more correlated with accident severity, than not wearing one. Likewise, from Figure 8, it was noted that persons involved in accidents aged 65 and above were more likely to be involved in more severe accidents, as is evidenced by the mutual information of 0.0022 in Figure 8. The node types, including non-intersections such as freeways and intersections, are also one of the more correlated features with accident severity, albeit with very similar mutual information of about 0.015. However, since direction is not specified, it could have contrasting effects on accident severity (see discussion on Figure 6).

For the supervised learning models, the three highest correlated features were selected to predict accident severity to alleviate the curse of dimensionality problem. Since there were numerous subfeatures from our predetermined factors, this would have rendered all data points far away from each other in very high dimensional space.

SLM Classification Models - Decision Tree and KNN for the three highest-correlated factors

As part of the investigation as to whether well-known predetermined features can predict accident severity, it is essential to utilise our labelled datasets to implement supervised learning models. Since the investigation was a classification of the categorial, target variable, accident severity, linear regression would not be an appropriate supervised learning model choice. 'SEVERITY' represents distinct classes (Fatality, Serious Injury, Other Injury, Non-Injury). Instead, k-Nearest Neighbours (k-NN) and Decision Tree Classifiers were preferred as alternatives, as these models are better suited to classify using (largely) categorical one-hot encoded data. They are also capable of producing interpretable class labels.

For both classification models - Decision Tree ad k-NN, an 80/20 split was used for training and testing, to ensure a sufficient part of the data was available for model training. A random state was provided as a parameter to ensure reproducibility of visualisations and figures. This particular split also allowed for an unbiased hold-out set for final evaluations. The remaining 20% test set refers to unseen data, when assessing the model's generalisation. In k-NN, when tuning hyperparameters, a further split of the training set was done, to 80% training and 20% validation. This ensured the test data remained unaffected during model selection, and proper separation of model development and the final evaluation phases was maintained.

In the k-NN model, the values of k were varied from 2 to 13 (inclusive), and evaluated the predictive performance of the three-highest correlated predetermined factors from Figure 11 - age of the elderly as well as whether or not the node was an intersection or not. This evaluation was done using metrics, like accuracy, precision, recall and F1 scores on the validation set, allowing the investigation to select a value of k that efficiently balances bias and variance.

The curse of dimensionality affects models like Decision Trees and k-NN, as it is prevalent in high-dimensional datasets. This is because as dimensions increase, distance metrics grow less meaningful and predictive performance degrades. In addressing this, the investigation performed feature selection, before modelling, to retain the most relevant features. Feature selection was performed manually, prior to fitting the models, by specifying the subset of three features that had the highest correlation with the target variable - accident severity (Figure 11). Hence, feature selection allowed for improving model interpretability, while reducing noise and any overfitting risks.

Both the Decision Tree and K-Nearest Neighbours (KNN) models were applied to classify accident severity using the three most informative features identified using mutual information: elderly age group and node type. Both the models showed similar trends and limitations in the output. The Decision Tree achieved an accuracy of 63.2% and always predicted the majority class, 'Other Injury', for all instances. This resulted in a recall of 1.0 for this class but 0.0 for all others, indicating complete failure to identify serious or fatal cases. Similarly, the KNN model recorded a slightly lower accuracy of 62.5% but showed slightly improved weighted F1-score (0.519 vs 0.490), as it occasionally predicted 'Serious Injury' cases, although with low recall (0.058). Both models struggled due to class imbalance and the limited predictive strength of the input features. However, KNN demonstrated slightly greater sensitivity to minority classes compared to Decision Tree, which was dominated by the majority class. While the Decision Tree is advantageous in interpretability and speed, it simplified the data. However KNN, although slightly more nuanced in capturing different relationships, is more computationally expensive with longer runtimes. Overall, both models performed poorly in identifying high-severity accidents, however Decision Tree was a much more efficient and faster classification model.

Limitations and improvement opportunities

The factor, seatbelt use, resulted in 27% of the data being missing, when preprocessed. A potential reason for such a high missing proportion value, may be due to seatbelt use being a human factor, so it could be information that drivers/passengers may have withheld. If this column was handled via imputing mode, it would heavily skew the data and bias it towards the mode value, so imputation was not performed for this specific factor. While this was done to limit systematic bias, it may have significantly affected the results, since 27% of the seatbelt use data was 'Unknown.' Consequently, the factor, vehicle type, was retained in its original column form without mode imputation, since it had a 0% missing proportion value after preprocessing. This decision may have reduced the statistical accuracy and reliability of associations between the target variable (accident severity) and vehicle type. Lack of imputation, primarily of seatbelt use, may have caused analysis of a reduced or non-representative subset of the data, so residual bias may have persisted. The retaining of 'Unknown' categories introduced ambiguity and may have created misleading groupings in the classification models. Potential improvements could include collecting complete data, dropping the accident rows with 'Unknown' values completely, or changing the schema of the investigation to not handle missing values at all. The latter approach ensures consistency between all predetermined factors.

There are key limitations to one-hot encoding, including the increased dimensionality as a result of having k columns for each predetermined feature, instead of one for each categorical feature encoded using this technique. 'The Curse of Dimensionality' is a potential issue as all data points will seem to be far from each other in very high dimensional space, especially if many features are one-hot encoded. Furthermore, it may not be possible to compute the exact correlations between the predetermined features and accident severity, as the features will be split into several columns.

When merging, it was noticed that some tables had different numbers of rows, despite every dataset having the index key, ACCIDENT_NO. Discrepancies in row counts could indicate unmatched or missing entries, which may lead to an incomplete representation of certain accidents. To preserve the original data structure and avoid introducing artificial bias, these inconsistencies wereC not forcibly resolved. Instead, the inconsistency is included in the report as an observation, and a potential confounding factor in the analysis.

Another key limitation is that correlation only measures the association between two variables, and does not infer whether or not one variable caused a change in the other due to lingering confounding variables. This limitation cannot be improved within this investigation since the data provided was of vehicle accidents, clearly an observational experiment where there is no control to randomise over conditions. Furthermore, mutual information scores are always positive, and hence, do not guide us as to the direction of correlation between the feature and the target varia1ble. Pearson's correlation coefficient, which does guide as to the direction of correlation, is only applicable for linear relationships and continuous numerical data, not for discrete numerical data as this investigation includes. However, the direction of correlation can partially be inferred from the heat maps discussed.

When adopting the Supervised Learning Model (SLM), and comparing the classification models of Decision Tree and k-NN performed well, in terms of classification metrics. However, the k-NN classifier had an overly excessive runtime. This time consumption affected the efficiency of the investigation severely. Future analyses can improve this by implementing a more optimised or systematic approach for executing the k-NN model, potentially incorporating parallel processing, dimensionality reduction (PCA), or algorithmic enhancements to reduce computational cost. These improvements would enhance the scalability and practicality of k-NN in larger or more complex datasets.

Conclusion

This investigation explored the influence of selected human, road and vehicle-related factors on accident severity outcomes in Victoria. By utilising the Victorian Road Crash dataset provided to us and applying data preprocessing techniques, such as imputation, one-hot encoding, and discretisation, the data was transferred into a format suitable for analysis and machine learning. Through a combination of mutual information scoring and visual analysis using heatmaps, multiple patterns were found that provided insight into how different conditions correlate with accident outcomes.

Human factors such as seatbelt use and age group, especially the elderly (65+), showed the strongest associations with accident severity. Road conditions such as road surface conditions and atmospheric factors also contributed, although with lower predictive power. In the context of vehicle factors, heavy freight vehicles such as Prime Movers and Rigid Trucks were associated with a higher proportion of severe accidents, while vehicle types such as cars and light commercial vehicles had comparatively lower severity outcomes.

However, despite the application of classification models like Decision Trees and K-Nearest Neighbours, the predictive performance of these models was limited. Notably, the models were skewed toward predicting the majority class (‘Other Injury’), highlighting the challenges posed by class imbalance and limited feature correlation strength.

Ultimately, to conclude the investigation, it was found that the chosen predetermined features are very insignificant in determining accident severity. However, there were many limitations to this analysis as already discussed as well as the investigation being based on accident data from only Victorian roads. If these limitations were improved or if there was better quality data available, a similar investigation may lead to contrasting conclusions.

Overall, this analysis has substantial real-world value. Identifying that factors such as seatbelt use, elderly age groups, and heavy vehicle types are consistently associated with more severe outcomes can inform targeted interventions. This can include public safety campaigns focused on seat belt compliance, tailored infrastructure for high-risk areas, and regulatory measures for heavy vehicle operation. This data-driven approach offers a way to implement effective strategies to reduce the severity of road accidents in Victoria.

References

Buuren, S. (n.d.). https://stefvanbuuren.name/fimd/. In *stefvanbuuren.name*. CRC Press. https://stefvanbuuren.name/fimd/

chugh, aakarsha. (2018, October 15). *ML | Label Encoding of datasets in Python*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/

Ganji, L. (2019, June 12). *One Hot Encoding in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-one-hot-encoding/

GeeksforGeeks. (2021, December 16). *How to Exclude Columns in Pandas?* GeeksforGeeks. https://www.geeksforgeeks.org/how-to-exclude-columns-in-pandas/

in. (2023, August 19). *How to Select Rows from a DataFrame Based on List Values in a Column in Pandas | Saturn Cloud Blog*. Saturncloud.io. https://saturncloud.io/blog/how-to-select-rows-from-a-dataframe-based-on-list-values-in-a-column-in-pandas/

Victoria Road Crash Data - Victorian Government Data Directory. (2024, December 11). Discover.data.vic.gov.au. https://discover.data.vic.gov.au/dataset/victoria-road-crash-data

W3Schools. (n.d.). *Matplotlib Pie Charts*. Www.w3schools.com. https://www.w3schools.com/python/matplotlib_pie_charts.asp